Quality control of FAstQ files using fastQC

fastqc ../Data/DNAseq/\*

Trim poor quality reads with TRim galore

trim\_galore --phred64 --fastqc -o ../LmexTG --paired -q 20 Data/DNAseq/LmexWT\_1.fastq Data/DNAseq/LmexWT\_2.fastq

trim\_galore --phred64 --fastqc -o ~/LmexTG/ --paired -q 20 Data/DNAseq/LmexAmpB\_1.fastq Data/DNAseq/LmexAmpB\_2.fastq

Re-run Fastqc on output

Align reads to reference genome using Bowtie

bowtie2 -p 2 -q -S WTout -x Lmex -1 LmexWT\_1\_val\_1.fq -2 LmexWT\_2\_val\_2.fq &

bowtie2 -p 2 -q -S AmpBout -x Lmex -1 LmexAmpB\_1\_val\_1.fq -2 LmexAmpB\_2\_val\_2.fq &

Sorting and removing duplicates.

samtools view -buS WTout.sam | samtools sort -o - WTout | samtools rmdup - WT

Assess coverage

genomeCoverageBed -ibam LmexWTpic.bam -g ~/Data/Reference/TriTrypDB-25\_LmexicanaMHOMGT2001U1103.fa > coverageWT.txt

This creates a VCF file called filtered\_snps.vcf, containing all the original SNPs from the raw\_snps.vcf file, but now the SNPs are annotated with either PASS or FILTER depending on whether or not they passed the filters.

java -Xmx3g -jar /usr/local/bin/GenomeAnalysisTK.jar -T VariantFiltration -R ~/Data/Reference/TriTrypDB-25\_LmexicanaMHOMGT2001U1103.fa -V AmpB.snps.raw.vcf -o AmpB.snps.filt.vcf --clusterSize 3 --clusterWindowSize 10 --filterExpression "QD < 2.0" --filterName "QDFilter" --filterExpression "MQ < 40.0" --filterName "MQFilter" --filterExpression "FS > 60.0" --filterName "FSFilter" --filterExpression "MQRankSum < -12.5" --filterName "MQRankSumFilter" --filterExpression "ReadPosRankSum < -8.0" --filterName "ReadPosFilter" --filterExpression "MQ0 >= 4 && ((MQ0 / (1.0 \* DP)) > 0.1)" --filterName "HARD\_TO\_VALIDATE" --filterExpression "QUAL < 30.0 || DP < 6 || DP > 5000" --filterName "QualFilter" -l INFO

java -Xmx3g -jar /usr/local/bin/GenomeAnalysisTK.jar -T VariantFiltration -R ~/Data/Reference/TriTrypDB-25\_LmexicanaMHOMGT2001U1103.fa -V AmpB.snps.raw.vcf -o AmpB.snps.filt.vcf --clusterSize 3 --clusterWindowSize 10 --filterExpression "QD < 2.0" --filterName "QDFilter" --filterExpression "MQ < 40.0" --filterName "MQFilter" --filterExpression "FS > 60.0" --filterName "FSFilter" --filterExpression "MQRankSum < -12.5" --filterName "MQRankSumFilter" --filterExpression "ReadPosRankSum < -8.0" --filterName "ReadPosFilter" --filterExpression "MQ0 >= 4 && ((MQ0 / (1.0 \* DP)) > 0.1)" --filterName "HARD\_TO\_VALIDATE" --filterExpression "QUAL < 30.0 || DP < 6 || DP > 5000" --filterName "QualFilter" -l INFO

java -Xmx3g -jar /usr/local/bin/GenomeAnalysisTK.jar -T VariantFiltration -R ~/Data/Reference/TriTrypDB-25\_LmexicanaMHOMGT2001U1103.fa -V AmpB.indel.raw.vcf -o AmpB.indel.filt.vcf --filterExpression "QD < 2.0" --filterName "QDFilter" --filterExpression "ReadPosRankSum < -20.0" --filterName "ReadPosFilter" --filterExpression "FS > 200.0" --filterName "FSFilter" --filterExpression "MQ0 >= 4 && ((MQ0 / (1.0 \* DP)) > 0.1)" --filterName "HARD\_TO\_VALIDATE" --filterExpression "QUAL < 30.0 || DP < 6 || DP > 5000" --filterName "QualFilter" -l INFO

java -Xmx3g -jar /usr/local/bin/picard.jar AddOrReplaceReadGroups I=LmexWT.bam O=LmexWTpic.bam RGID=1 RGLB=rgLib RGPL=illumina RGPU=single RGSM=WT

java -Xmx3g -jar /usr/local/bin/picard.jar AddOrReplaceReadGroups I=LmexAmpB.bam O=LmexAmpBpic.bam RGID=1 RGLB=rgLib RGPL=illumina RGPU=single RGSM=Resistant

samtools view -H LmexAmpBpic.bam # look at header only

samtools index LmexAmpBpic.bam

samtools index LmexWTpic.bam

Check genome coverage

genomeCoverageBed -ibam LmexWTpic.bam -g ~/Data/Reference/TriTrypDB-25\_LmexicanaMHOMGT2001U1103.fa > coverageWT.txt

genomeCoverageBed -ibam LmexAmpBpic.bam -g ~/Data/Reference/TriTrypDB-25\_LmexicanaMHOMGT2001U1103.fa > coverageWT.txt

java -Xmx3g -jar /usr/local/bin/picard.jar CreateSequenceDictionary R=~/Data/Reference/TriTrypDB-25\_LmexicanaMHOMGT2001U1103.fa O=Lmex.dict

samtools faidx ~/Data/Reference/TriTrypDB-25\_LmexicanaMHOMGT2001U1103.fa

samtools faidx ~/Data/Reference/TriTrypDB-25\_LmexicanaMHOMGT2001U1103.fa

Call variants on a haploid genome with the UnifiedGenotyper, producing a raw (unfiltered) VCF.

java -Xmx3g -jar /usr/local/bin/GenomeAnalysisTK.jar -T UnifiedGenotyper -R ~/Data/Reference/TriTrypDB-25\_LmexicanaMHOMGT2001U1103.fa -I LmexAmpBpic.bam -o AmpB.snps.raw.vcf -stand\_call\_conf 30.0 -stand\_emit\_conf 10.0 -out\_mode EMIT\_VARIANTS\_ONLY -l INFO -A Coverage -A HaplotypeScore -A InbreedingCoeff -glm SNP -nt 2

This creates a VCF file called filtered\_snps.vcf, containing all the original SNPs from the raw\_snps.vcf file, but now the SNPs are annotated with either PASS or FILTER depending on whether or not they passed the filters.

java -Xmx3g -jar /usr/local/bin/GenomeAnalysisTK.jar -T VariantFiltration -R ~/Data/Reference/TriTrypDB-25\_LmexicanaMHOMGT2001U1103.fa -V AmpB.snps.raw.vcf -o AmpB.snps.filt.vcf --clusterSize 3 --clusterWindowSize 10 --filterExpression "QD < 2.0" --filterName "QDFilter" --filterExpression "MQ < 40.0" --filterName "MQFilter" --filterExpression "FS > 60.0" --filterName "FSFilter" --filterExpression "MQRankSum < -12.5" --filterName "MQRankSumFilter" --filterExpression "ReadPosRankSum < -8.0" --filterName "ReadPosFilter" --filterExpression "MQ0 >= 4 && ((MQ0 / (1.0 \* DP)) > 0.1)" --filterName "HARD\_TO\_VALIDATE" --filterExpression "QUAL < 30.0 || DP < 6 || DP > 5000" --filterName "QualFilter" -l INFO

java -Xmx3g -jar /usr/local/bin/GenomeAnalysisTK.jar -T VariantFiltration -R ~/Data/Reference/TriTrypDB-25\_LmexicanaMHOMGT2001U1103.fa -V AmpB.snps.raw.vcf -o AmpB.snps.filt.vcf --clusterSize 3 --clusterWindowSize 10 --filterExpression "QD < 2.0" --filterName "QDFilter" --filterExpression "MQ < 40.0" --filterName "MQFilter" --filterExpression "FS > 60.0" --filterName "FSFilter" --filterExpression "MQRankSum < -12.5" --filterName "MQRankSumFilter" --filterExpression "ReadPosRankSum < -8.0" --filterName "ReadPosFilter" --filterExpression "MQ0 >= 4 && ((MQ0 / (1.0 \* DP)) > 0.1)" --filterName "HARD\_TO\_VALIDATE" --filterExpression "QUAL < 30.0 || DP < 6 || DP > 5000" --filterName "QualFilter" -l INFO

java -Xmx3g -jar /usr/local/bin/GenomeAnalysisTK.jar -T VariantFiltration -R ~/Data/Reference/TriTrypDB-25\_LmexicanaMHOMGT2001U1103.fa -V AmpB.indel.raw.vcf -o AmpB.indel.filt.vcf --filterExpression "QD < 2.0" --filterName "QDFilter" --filterExpression "ReadPosRankSum < -20.0" --filterName "ReadPosFilter" --filterExpression "FS > 200.0" --filterName "FSFilter" --filterExpression "MQ0 >= 4 && ((MQ0 / (1.0 \* DP)) > 0.1)" --filterName "HARD\_TO\_VALIDATE" --filterExpression "QUAL < 30.0 || DP < 6 || DP > 5000" --filterName "QualFilter" -l INFO

SNPEff to annotate the VCF files . SnpEff (Version: 4.1 (2015-01))a variant annotation and effect prediction tool. It annotates and predicts the effects of genetic variants (such as amino acid changes). The inputs are predicted variants, It annotates the variants and calculates the effects they produce on known genes (e.g. amino acid changes)

java -Xmx3g -jar /usr/local/bin/snpEff.jar TriTrypDB-25\_LmexicanaMHOMGT2001U1103 ~/Data/LmexWT.snps.filt.vcf > WT.snps.ann.vcf

Output variants that cause a codon to produces a different amino acid (missense\_variant)

cat WT.snps.ann.vcf | java -Xmx3g -jar /usr/local/bin/SnpSift.jar filter "ANN[\*].EFFECT has 'missense\_variant'" > WT.snps.sift.vcf

cat AmpB.snps.ann.vcf | java -Xmx3g -jar /usr/local/bin/SnpSift.jar filter "ANN[\*].EFFECT has 'missense\_variant'" > AmpB.snps.sift.vcf

vcfTools

vcftools --vcf ~/Data/WT.snps.sift.vcf --diff ~/Data/AmpB.snps.sift.vcf --out compare --diff-site

SnpSift for filtered data

cat WT.snps.sift.vcf | java -Xmx3g -jar /usr/local/bin/SnpSift.jar filter "(FILTER='PASS')" > WT.snps.pass.vcf

cat AmpB.snps.sift.vcf | java -Xmx3g -jar /usr/local/bin/SnpSift.jar filter "(FILTER='PASS')" > AmpB.snps.pass.vcf

vcfTools for filtered data (vcftools v0.1.14) Outputs the sites that are common / unique to each file.

vcftools --vcf ~/Data/WT.snps.pass.vcf --diff ~/Data/AmpB.snps.pass.vcf --out comparePassed --diff-site

grep -f e

grep -e 'LmxM.11.1100' ergosterol\_WT.txt | grep -e '443299'

vcfTools for for ergosterol pathway

vcftools --vcf ~/pathway/ergosterol\_WT.txt --diff ~/Data/ergosterol\_AmpB.txt --out comparePathway --diff-site

Matthews Scripts

ubuntu@ip-172-31-36-102:~/pathway$ awk '{ print "\\^"$1"."$3 }' ampBsnps.txt | while read line; do grep $line ../Data/AmpB.snps.pass.vcf; done > AmpB.snps.unique.vcf

ubuntu@ip-172-31-36-102:~/pathway$ awk '{ print $1 }' GenesByMetabolicPathwayKegg\_summary.txt | grep -e - AmpB.snps.unique.vcf

LmxM.11 443299 . A T 2304.77 PASS AC=2;AF=1.00;AN=2;DP=64;Dels=0.00;FS=0.000;HaplotypeScore=2.2880;MLEAC=2;MLEAF=1.00;MQ=44.00;MQ0=0;QD=30.59;SOR=0.822;ANN=T|missense\_variant|MODERATE|LmxM.11.1100|LmxM.11.1100|transcript|rna\_LmxM.11.1100-1|Coding|1/1|c.527A>T|p.Asn176Ile|527/1440|527/1440|176/479||,T|upstream\_gene\_variant|MODIFIER|LmxM.11.1130|LmxM.11.1130|transcript|rna\_LmxM.11.1130-1|Coding||c.-3635A>T|||||3635|,T|downstream\_gene\_variant|MODIFIER|LmxM.11.1090|LmxM.11.1090|transcript|rna\_LmxM.11.109 -1|Coding||c.\*2771A>T|||||2771| GT:AD:DP:GQ:PL 1/1:0,64:64:99:2333,183,0

ubuntu@ip-172-31-36-102:~/pathway$

my script

awk '{print $1,"$2 }' ~/SNPs1\_WT.txt |grep '$1 AmpB.snps.pass.vcf| grep -e $2 }'